На правах рукописи

Ковалевский Артем Павлович

СТАТИСТИЧЕСКИЕ КРИТЕРИИ АПОСТЕРИОРНОГО ОБНАРУЖЕНИЯ РАЗЛАДКИ ВРЕМЕННЫХ РЯДОВ И ИХ ПРИМЕНЕНИЯ

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
доктора физико-математических наук

Новосибирск — 2018

Работа выполнена в Новосибирском государственном техническом университете

Официальные оппоненты:

Войтишек Антон Вацлавович, доктор физико-математических наук, профессор, лаборатория стохастических задач Института вычислительной математики и математической геофизики Сибирского отделения Российской академии наук, ведущий научный сотрудник

Лотов Владимир Иванович, доктор физико-математических наук, профессор, лаборатория теории вероятностей и математической статистики Института математики Сибирского отделения Российской академии наук, заведующий лабораторией

Поддубный Василий Васильевич, доктор технических наук, профессор, кафедра прикладной информатики, лаборатория когнитивных исследований языка Национального исследовательского Томского государственного университета, профессор, главный научный сотрудник

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт прикладной математики Дальневосточного отделения Российской академии наук

Защита состоится:

21 марта 2019 г. в 15.00 на заседании диссертационного совета Д 999.082.03 на базе Федерального государственного бюджетного учреждения науки Института систем информатики им. А.П. Ершова Сибирского отделения Российской академии наук (ИСИ СО РАН) по адресу 630090, г. Новосибирск, проспект Академика Лаврентьева, 6, комн. 254.

С диссертацией можно ознакомиться в библиотеке и на сайте ИСИ СО РАН http://www.iis.nsk.su/files/Kovalevskiy_dis.pdf

Автореферат разослан: "_____ "_____ 2018 г.

Ученый секретарь
диссертационного совета
канд. физ.-мат. наук  Мурзин Федор Александрович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. В диссертации исследуется задача проверки однородности временного ряда. В простейшей постановке наблюдается временной ряд, для которого рассматриваются две вероятностные модели: согласно основной гипотезе, элементы временного ряда являются независимыми одинаково распределенными случайными величинами; согласно альтернативной гипотезе, в некоторый момент внутри интервала наблюдения происходит разладка: случайные величины по-прежнему предполагаются независимыми, но до момента разладки имеют одно распределение, а после момента разладки — другое. Задача оценивания параметров разладки и изучения вероятностных свойств оценок решалась А. А. Боровковым и Ю.Ю. Линке, Вальдом (A. Wald), И. В. Никифоровым, Н. Клигене и Л. Телькснисом, И. Ш. Торговицким, А. Н. Ширяевым, В. И. Лотовым, Карлстейном (E. Carlstein), Ксорго и Хорвафом (M. Csorgo, L. Horvath), Дембгеном (L. Dümbgen), Шабаном (S. A. Shaban).

Различают две задачи обнаружения разладки: последовательную процедуру и апостериорную. При последовательной процедуре значения появляются одно за другим, и акцент делается на наискорейшем обнаружении разладки при фиксированных вероятностях ошибок. Для обнаружения разладки используется последовательный критерий отношения правдоподобия Вальда. А. Н. Ширяевым изучены последовательные решающие правила для марковских процессов. В работе С. Э. Воробейчикова и Ю. С. Пономаревой предлагается применение последовательного метода наименьших квадратов для обнаружения разладки авторегрессионных моделей и их обобщений.

Задача апостериорного обнаружения разладки состоит в том, что временной ряд известен полностью, и надо сделать выбор между моделью выборки и моделью разладки. Эта задача рассмотрена в книге Б. Е. Бродского и Б. С. Дарховского. Отметим, что как в этой книге, так и в монографии Ксорго и Хорвафа отсутствует сравнение критериев обнаружения разладки на основании относительной асимптотической эффективности по Питмену. Между тем такое сравнение, как будет показано ниже, приводит к выбору единственного критерия, относительно наиболее асимптотически эффективного по Питмену в широком классе, и потому наиболее подходящего для практического различения близких гипотез.

Математические подходы к анализу текстов развиты в работах Н. А. Морозова, А. А. Маркова, Г. Хетсо, Д. В. Хмелева, А. А. Поликарпова, В. В. Поддубного и О. Г. Шевелева. Применение разработанных в диссертации критериев к анализу однородности художественных текстов на русском языке осуществляется на основании метода, предложенного Н. А. Морозовым и развитого рядом авторов, в том числе В. П. Фоменко и Т. Г. Фоменко. Метод состоит в том, что по тексту строится временной ряд индикаторов служебных слов (предлогов, союзов, частиц): выбирается значение 1, если слово является служебным, и значение 0 иначе. Обобщением этого метода является метод построения нескольких связанных временных рядов по тексту. Временные ряды могут отражать, в частности, число букв или слогов в слове, число слов в предложении (Г. Хетсо). Однако все исследованные нами временные ряды, за исключением ряда индикаторов служебных слов, не обладают в нужной степени селективностью автора: методами дисперсионного анализа можно убедиться в том, что внутригрупповые дисперсии

для текстов каждого автора обеспечивают основной вклад в суммарную дисперсию, и различия в средних значениях характеристик для разных авторов незначимы. Таким образом, эти характеристики (число букв в слове и т.п.) менее полезны для анализа однородности текста.

Результат применения статистического критерия к анализу однородности каждого конкретного текста — достигнутый уровень значимости гипотезы об однородности. Чем меньше (ближе к нулю) достигнутый уровень значимости для данного текста, тем более это говорит против гипотезы об однородности. Как правило, для более длинных текстов одного автора достигнутые уровни значимости ниже, чем для коротких. Но значительно ниже достигнутые уровни значимости для текстов, полученных склейкой (конкатенацией) двух произведений разных авторов. Разработанный метод позволяет не только диагностировать наличие разладки, то есть неприемлемость модели выборки, но и указывать момент разладки — склейка текстов, содержащих многие сотни страниц, отыскивается с точностью порядка одной страницы.

Однако отмеченное выше свойство уменьшения достигаемого уровня значимости с ростом объема текста говорит о том, что модель выборки не является удовлетворительной для временного ряда, полученного по произведению или собранию сочинений автора. Действительно, статистический анализ показывает наличие корреляций, медленно убывающих с ростом лага. Подходящей моделью для таких временных рядов является модель фрактального шума.

Фрактальный гауссовский шум — это стационарная гауссовская последовательность с нулевым математическим ожиданием и корреляционной функцией, убывающей по степенному закону таким образом, что для частичных сумм S_n выполнено равенство $\mathbf{D}S_n = \sigma^2 n^{2H}$, где $H \in (0,\ 1]$ называется показателем Херста. В случае склейки текстов двух авторов получается модель разладки фрактального гауссовского шума.

Модель фрактального гауссовского шума введена независимо друг от друга А. Н. Колмогоровым и Винером в 1940 году, а свое название и широкую известность получила благодаря статье Мандельброта и ван Несса. Она оказывается полезной для описания экономических и естественно-научных временных рядов. Использование модели осложнено трудоемкой процедурой оценивания ее параметров.

Для решения этой проблемы автором разработан бинарный знаковый метод оценивания параметра H, имеющий низкую вычислительную сложность. Дисперсия оценки, полученной этим методом, не намного больше дисперсии оптимальной оценки. Асимптотика дисперсии оценки вычислена аналитически. Разработаны и реализованы алгоритмы моделирования фрактального гауссовского шума и алгоритм оценивания параметра Херста бинарным знаковым методом. Результаты моделирования согласуются с результатами, полученными аналитическими методами теории вероятностей.

Бинарный знаковый метод оценивания параметра H применяется к временным рядам, построенным по текстам на естественном языке с помощью авторского инварианта. На его основании строится статистический критерий проверки гипотезы об отсутствии фрактальности. Критерий основан на разности между оценкой параметра Херста и значением $1/2$, соответствующим белому гауссовскому шуму. Построенный критерий позволяет при изучении текстов обосновать выбор модели фрактального шума про-

тив альтернативной модели выборки. Этот вывод делается на основании того, что коэффициент Херста значимо отличается от 1/2.

Для исследования разладки в текстах, то есть для выбора между моделью фрактального гауссовского шума и разладки фрактального гауссовского шума, разработан статистический критерий, основанный на разности оценок разными методами: методом дисперсии и бинарным знаковым методом. Этот критерий дает хорошие результаты при анализе однородности текста: достигнутые уровни значимости далеки от нуля для текстов одного автора и близки к нулю для склейки текстов разных авторов.

Обнаружение разладки процессов гармонических колебаний со случайным шумом (возникающих, в частности, при анализе колебаний инженерных конструкций) основано на изучении асимптотического поведения сумм остатков соответствующей регрессионной модели. Это изучение было осуществлено МакНилом в 1978 году. В диссертации доказана теорема, распространяющая результат МакНила на эмпирический мост. Получены следствия о предельных распределениях ряда функционалов от эмпирического моста. Сравнение статистических критериев, основанных на этих функционалах, проведено на численном примере в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени. Сравнение проводилось путем моделирования и выявило преимущество функционала супремума отклонения эмпирического моста.

Цель работы

Целью работы является разработка статистических критериев проверки гипотез для класса вероятностных моделей временных рядов. Этот класс включает в себя модели случайной выборки и ее разладки, фрактального гауссовского шума и его разладки, циклического тренда со случайным шумом и его разладки. Разработанные критерии апостериорного обнаружения разладки применяются для анализа однородности текстов и для выбора адекватной вероятностной модели в ряде задач анализа медицинских и экономических данных.

В рамках указанной цели были поставлены следующие задачи.

1. Построить класс статистических критериев проверки однородности временного ряда, основанных на функционалах от эмпирического моста. Сравнить критерии из этого класса в смысле асимптотической относительной эффективности по Питмену и выбрать наилучший критерий.

2. Построить класс статистических критериев проверки однородности временного ряда, основанных на предположениях нормальности.

3. Разработать знаковый метод оценивания параметра Херста и его модификации для моделей фрактального гауссовского шума и фрактального броуновского моста.

4. Построить класс статистических критериев для различения модели выборки и фрактального шума; модели фрактального шума и его разладки.

5. Применить разработанные критерии обнаружения разладки к анализу однородности текста. Формализовать модели временных рядов, по-

строенных по тексту одного автора и по склейке текстов разных авторов. Исследовать результативность применения разработанных статистических критериев проверки однородности к текстам на естественном языке и обосновать алгоритм выявления склейки текстов.

6. Применить разработанные критерии для выбора адекватной вероятностной модели к анализу медицинских и экономических данных.

7. Построить статистические критерии обнаружения разладки процесса гармонических колебаний со случайным шумом, пригодные для использования при компьютерном анализе изменений прочностных характеристик конструкции на основании записи ее колебаний.

8. На численном примере сравнить предложенные критерии в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени, равномерно распределенный на интервале наблюдения.

Методы исследования. Исследования, проведенные в работе, основаны на применении и развитии методов статистического анализа временных рядов, в частности, методов апостериорного обнаружения разладки, изложенных в монографиях Б. Е. Бродского и Б. С. Дарховского, Ксорго и Хорвафа; методов сравнения критериев, изложенных в книге Я. Ю. Никитина; знакового метода оценивания корреляционной функции А. М. Яглома; алгоритмов оценивания параметров и проверки гипотез для фрактального гауссовского шума, развитых в ряде работ последних десятилетий; математических методов классификации текстов, развитых в работах Н. А. Морозова, А. А. Маркова, Г. Хетсо, Д. В. Хмелева, В. В. Поддубного и О. Г. Шевелева; классических методов доказательства предельных теорем теории вероятностей в функциональных пространствах.

Апробация работы

Основные результаты диссертации докладывались и обсуждались на следующих всероссийских и международных конференциях:

- Workshop on Mathematics of Stochastic Networks, EURANDOM, The Netherlands, November 2001.
- 6th International Symposium on science and technology (KORUS–2002), Novosibirsk, June 24–26, 2002.
- Квантитативная лингвистика: исследования и модели (КЛИМ–2005), Новосибирский государственный педагогический университет, Новосибирск, 6–10 июня 2005 г.
- IV International Conference on Limit Theorems in Probability Theory and Their Applications, Novosibirsk, 21–25 August 2006.
- XIV Всероссийская школа-коллоквиум по стохастическим методам и VIII Всероссийский симпозиум по прикладной и промышленной математике, Сочи-Адлер, 29 сентября – 7 октября 2007 г.

- IX Международная научно-техническая конференция «Актуальные проблемы электронного приборостроения АПЭП-2008», Новосибирский Государственный Технический Университет, г. Новосибирск, 24–26 сентября 2008 г.
- Programme «Stochastic Processes in Communication Sciences», Isaac Newton Institute for Mathematical Sciences, Cambridge, 25 May – 15 June 2010.
- X Международная научно-техническая конференция «Актуальные проблемы электронного приборостроения АПЭП-2010», Новосибирский государственный технический университет, г. Новосибирск, 22–24 сентября 2010 г.
- III Всероссийский семинар «Фундаментальные основы МЭМС- и нанотехнологий». Новосибирск, 25–27 мая 2011 г.
- V International Conference «Limit Theorems in Probability Theory and Their Applications». Novosibirsk, August 15–21, 2011.
- 22nd Annual Conference of The International Environmetrics Society. Hyderabad, India, January 1–6, 2012.
- Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control. Novosibirsk, September 26–30, 2013.
- 11 International conference on ordered statistical data. Bedlewo, Poland, June 2–6, 2014.
- XXI Всероссийская школа-коллоквиум по стохастическим методам. Кисловодск, 11–17 июня 2017 г.
- 13 International conference on ordered statistical data. Cadiz, Spain, May 22–25, 2018.

Кроме того, основные результаты диссертации докладывались неоднократно на:

- объединенном семинаре кафедры теории вероятностей и математической статистики НГУ и лаборатории теории вероятностей и математической статистики ИМ СО РАН под руководством академика А. А. Боровкова;
- научном семинаре кафедры высшей математики Новосибирского государственного технического университета под руководством профессора В. А. Селезнева;
- научных сессиях факультета прикладной математики и информатики НГТУ под руководством профессора Б. Ю. Лемешко;
- научном семинаре кафедры вычислительной техники Новосибирского государственного технического университета под руководством профессора В. В. Губарева.

Публикации

По теме диссертации автором опубликовано 48 печатных работ, в том числе 11 работ, индексируемых в базах цитирования (RSCI, SCOPUS, WoS). 9 работ опубликовано без соавторов.

Личный вклад автора

Диссертационная работа выполнена непосредственно ее автором.

В совместных работах [12], [18], [13] автору диссертации принадлежат вывод расчетных формул, реализация расчетов и интерпретация их результатов; в работах [2], [5], [6] — постановка задачи, доказательство утверждений и разработка алгоритмов, интерпретация результатов расчетов; в работах [1], [15], [16] — постановка задачи и доказательство утверждений; в работах [19], [20], [29], [32] — вывод расчетных формул и разработка алгоритмов, интерпретация результатов расчетов; в работе [35] — постановка задачи и разработка вычислительных алгоритмов.

Работа выполнялась в Новосибирском государственном техническом университете в период с 1999 по 2018 год.

Структура и объем диссертации

Диссертация состоит из введения, шести глав и заключения. Главы делятся на параграфы. Нумерация утверждений двойная: например, теорема 2.1 является первой теоремой главы 2. Нумерация формул сквозная. Общий объем диссертации 271 страница, список литературы содержит 210 наименований.

Научная новизна

1. Впервые введен широкий класс статистических критериев апостериорного обнаружения разладки в модели выборки с единых позиций: критерии построены на основании функционалов от эмпирического моста. Рассмотрены функционалы, основанные на взвешенных суммах; L_p-нормы и их модификации; L_∞-норма и размах эмпирического моста.

2. Впервые проведено сравнение статистических критериев апостериорного обнаружения разладки в модели выборки с точки зрения их относительной асимптотической эффективности по Питмену. Выбран наилучший критерий, основанный на L_∞-норме эмпирического моста.

3. Впервые разработан алгоритм применения статистического критерия, основанного на норме эмпирического моста, к анализу однородности текста на естественном языке. Выявлена зависимость реально достигаемого уровня значимости критерия от объема текста, приводящая к неадекватности модели выборки для текста большого объема.

4. Впервые предложены модели фрактального броуновского моста и склейки фрактальных броуновских движений для временного ряда, построенного по тексту на естественном языке.

5. Впервые предложены центрированный знаковый метод, модифицированный знаковый метод и бинарный знаковый метод оценивания параметра Херста. Построен статистический критерий проверки гипотезы фрактальности.

6. Впервые статистический критерий проверки гипотезы фрактальности применен для проверки фрактальности текстов на естественном языке.

7. Впервые построен статистический критерий проверки разладки фрактального гауссовского шума, основанный на разности оценок параметра Херста. Критерий применен к анализу однородности текста на естественном языке. Разработан алгоритм выявления склейки текстов.

8. Впервые разработан класс критериев обнаружения разладки регрессии с циклическим трендом, основанных на значениях эмпирического моста. На численном примере проведено сравнение мощностей разработанных критериев.

Теоретическая значимость

1. Метод построения критериев наличия разладки на основании функционалов от эмпирического моста систематизирует процедуры построения таких критериев и может быть использован для исследования разладки в более сложных, в том числе регрессионных, моделях.

2. Процедура сравнения критериев наличия разладки и выявления лучшего в широком классе может быть применена к другим вероятностным моделям.

3. Разработанные модификации знаковых методов оценивания параметра Херста становятся классом методов, инвариантных относительно строго монотонных преобразований пространства значений.

4. Построены статистические модели временного ряда, образованного по тексту на естественном языке, в однородном и неоднородном случаях. Эти модели протестированы на адекватность методами математической статистики. Проведенный анализ позволяет выбирать правильную модель в зависимости от объема текста, а также различать однородные и неоднородные тексты.

5. Разработаны статистические критерии обнаружения разладки регрессии с циклическим трендом.

Практическая значимость

1. Разработаны критерии апостериорного обнаружения разладки в модели выборки, позволяющие решать задачи об однородности случайных последовательностей. Обоснован выбор критерия, имеющего наибольшую относительную асимптотическую эффективность по Питмену, и потому наиболее полезного при практическом различении близких гипотез, то есть в ситуации, когда математические ожидания до и после разладки различаются незначительно.

2. Разработаны статистические критерии, позволяющие по статистике перемен знаков оценивать параметр Херста. В частности, принимать или отвергать модель фрактального броуновского движения. Также разработаны критерии, позволяющие диагностировать разладку в модели фрактального броуновского движения.

3. Методика применения разработанных статистических критериев к анализу однородности текста на естественном языке позволяет анализировать тексты на однородность.

4. Предложенные статистические критерии позволяют тестировать наличие разладки регрессионных моделей с циклическим трендом. В частности, они применялись в качестве решающих правил для выявления наличия или отсутствия изменений прочностных характеристик высотных и уникальных зданий на основании записей их колебаний, выполненных НПО «Содис» (г. Москва).

5. Предложенные критерии позволяют обнаруживать разладку в регрессионных моделях. Так, они применялись для выбора модели зависимости концентрации маркеров в крови от массы тела; зависимости цены квартиры от ее характеристик; зависимости цены автомобиля от года выпуска.

Достоверность результатов диссертации подтверждается их совпадением в частных случаях с результатами расчетов, выполненных другими авторами и с помощью других методов. Теоретические результаты опубликованы в ведущих журналах, докладывались на крупных международных конференциях и представлены в их публикациях. Они известны в научном сообществе и цитируются в работах других авторов.

На защиту выносятся

1. Построение класса статистических критериев (решающих правил) и соответствующих им алгоритмов апостериорного обнаружения разладки в модели выборки на основании функционалов от эмпирического моста.

2. Сравнение алгоритмов апостериорного обнаружения разладки в модели выборки с точки зрения их относительной асимптотической эффективности по Питмену.

3. Алгоритм применения статистического критерия, основанного на норме эмпирического моста, к анализу однородности текста на естественном языке. Выявление зависимости реально достигаемого уровня значимости критерия от объема текста, приводящая к неадекватности модели выборки для текста большого объема.

4. Модели фрактального броуновского моста и склейки фрактальных броуновских движений для временного ряда, построенного по тексту на естественном языке.

5. Центрированный знаковый метод, модифицированный знаковый метод и бинарный знаковый метод оценивания параметра Херста. Построение решающего правила и алгоритма проверки гипотезы фрактальности.

6. Алгоритм проверки фрактальной модели для текстов на естественном языке с помощью специально разработанного статистического критерия.

7. Разработка алгоритма проверки разладки фрактального гауссовского шума, основанного на разности оценок параметра Херста, его применение к анализу однородности текста на естественном языке. Разработка алгоритма выявления склейки текстов.

8. Функциональная предельная теорема для числа разных элементов выборки и алгоритмы ее применения к анализу текстов.

9. Алгоритмы обнаружения разладки регрессии на порядковые статистики и регрессии с циклическим трендом, основанные на функционалах от эмпирического моста.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** приводятся краткий обзор литературы, постановка задач и основные результаты диссертации.

В **главе 1** в качестве исходной модели однородного временного ряда используется модель выборки. Для неоднородного — модель разладки. В целях изучения предельного поведения процессов предполагается, что объем выборки растет, и модели формализуются как схемы серий независимых в каждой серии случайных величин $X_1^{(n)},\ \ldots,\ X_n^{(n)},\ n \geq 1$.

Предполагается, что случайные величины заданы $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин.

Случайные величины $\xi_i^{(1)}$ имеют распределение $\mathcal{F}_1$ с математическим ожиданием m_1 и дисперсией $0 < {\sigma_1}^2 < \infty$.

Случайные величины $\xi_i^{(2)}$ имеют распределение $\mathcal{F}_2$ с математическим ожиданием m_2 и дисперсией $0 < {\sigma_2}^2 < \infty$.

Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^{n}$ задается следующим образом:

$X_i^{(n)} = \xi_i^{(1)}$ при $1 \leq i \leq [nT]$;

$X_i^{(n)} = \xi_i^{(2)}$ при $[nT]+1 \leq i \leq n$.

Здесь T — неизвестная константа, $0 < T < 1$.

Простыми гипотезами θ здесь являются тройки $\theta = (\mathcal{F}_1,\ \mathcal{F}_2,\ T)$, фиксирующие распределения до и после разладки, а также момент разладки. Предполагается, что все множество гипотез Θ представимо в виде объединения $\Theta = \Theta_0 \cup \Theta_1$, где $\Theta_0 = \{\theta :\ \mathcal{F}_1 = \mathcal{F}_2,\ m_1 \neq 0\}$ соответствует отсутствию разладки, а $\Theta_1 = \{\theta :\ m_1 \neq m_2\}$ наличию разладки. Задача состоит в построении критериев, различающих сложные гипотезы Θ_0 и Θ_1.

Критерий отвергает гипотезу Θ_0 в том и только том случае, когда $J_n \geq C$, где J_n — статистика, определяющая критерий.

Рассматриваются последовательности статистик $\{J_n\}$, которые при выполнении любой $\theta_0 \in \Theta_0$ сходятся по распределению к случайной величине J с функцией распределения F, причем F не зависит от конкретной гипотезы θ_0. *Приближенный бахадуровский наклон* $c(\theta)$ для последовательности статистик $\{J_n\}$ и гипотезы $\theta \in \Theta_1$ определяется равенством

$$c(\theta) = 2 \lim_{n\to\infty} (-n^{-1} \ln(1 - F(J_n))) \quad (\mathbf{P}_\theta - \text{п. н.}) \tag{1}$$

Это определение будет нами использоваться для сравнения критериев: один критерий *лучше* другого в смысле используемого подхода, если для всех $\theta \in \Theta_1$ значение $c(\theta)$ приближенного бахадуровского наклона для статистики первого критерия не меньше (и хотя бы для одного $\theta \in \Theta_1$ строго больше), чем для второго. Этот подход предложен Питменом для статистик, имеющих в пределе нормальное распределение. Для случая, когда предельное распределение абсолютно непрерывно, но отлично от нормального, модификация подхода Питмена введена Виэндом.

Строим статистики, являющиеся функционалами от *эмпирического моста* $Z_n = \{Z_n(t), 0 \leq t \leq 1\}$ — случайной ломаной, построенной по точкам

$$\left(\frac{k}{n};\ \frac{nS_k - kS_n}{sn\sqrt{n}}\right), k = 0, \ldots, n, \tag{2}$$

где $S_n = \sum_{i=1}^n X_i^{(n)} = n\overline{X}$, $s^2 = \overline{X^2} - (\overline{X})^2$.

Рассматриваем различные нормы случайной функции Z_n на отрезке $[0;\ 1]$, а также размах этой функции на отрезке. Обозначим

$$J_n^{(r)} = ||Z_n||_{L_r} = \left(\int_0^1 |Z_n(t)|^r dt \right)^{1/r};$$

$$J_n^{\infty} = ||Z_n||_{L_\infty} = \sup_{t \in [0;\ 1]} |Z_n(t)|; \qquad J_n^R = \sup_{t \in [0;\ 1]} Z_n(t) - \inf_{t \in [0;\ 1]} Z_n(t).$$

Наряду с L_r-нормами эмпирического процесса рассматриваются функционалы вида

$$J_n^{|r|} = \left| \int_0^1 (Z_n(t))^r dt \right|^{1/r}. \tag{3}$$

При четном r эти функционалы совпадают с L_r-нормами.

Следующая теорема обосновывает неоптимальность использования ряда статистик (в частности, размаха эмпирического моста) для апостериорного обнаружения разладки.

Теорема 1.1 *Статистика J_n^R не лучше статистики J_n^{∞}. Статистика $J_n^{(r)}$ не лучше статистики $J_n^{|r|}$.*

Наряду с уже введенными рассмотрим статистики, основанные на взвешенных суммах элементов временного ряда:

$$J_n = \left| \frac{\sum_{k=1}^n h_{k,n} X_k}{s\sqrt{n}} \right|.$$

Здесь $h_{1,n}, \ldots, h_{k,n}$ — весовые коэффициенты. Мы будем предполагать, что весовые коэффициенты заданы следующим регулярным образом: $h_{k,n} = g(k/n)$, где $g(t)$ — функция ограниченной вариации на $[0,\ 1]$. Для введенных с помощью функции g весовых коэффициентов будем использовать обозначение статистики $J_n = J_n(g)$.

Следующая теорема дает необходимые и достаточные условия сходимости статистики $J_n(g)$ по распределению к модулю стандартного нормального закона.

Теорема 1.2 *Для любой $\theta_0 \in \Theta_0$ статистика $J_n(g)$ сходится по распределению к модулю стандартного нормального закона в том и только том случае, когда выполнены условия*

$$\int_0^1 g(t)dt = 0, \qquad \int_0^1 g^2(t)dt = 1. \tag{4}$$

При выполнении условий этой теоремы статистику $J_n = J_n(g)$ можно представить с помощью интеграла Стилтьеса.

Теорема 1.3 *Пусть g — функция ограниченной вариации, удовлетворяющая условиям (4) теоремы 1.2. Тогда статистику $J_n(g)$ можно представить в виде интеграла Стилтьеса*

$$J_n(g) = \left| \int_0^1 Z_n(t) dg(t) + \nu_n \right|,$$

где $\nu_n \to 0$ ($\mathbf{P}_\theta$-п.н.) для любой $\theta \in \Theta$.

В качестве следствия получаем усиленный закон больших чисел для этой статистики.

Для сравнения статистик взвешенных сумм рассмотрим байесовскую постановку задачи — будем предполагать, что T — случайная величина с известной функцией распределения $F_T(t)$.

В теореме 1.4 найден наилучший в классе критериев, удовлетворяющих условиям теоремы 1.3 в ситуации, когда известна функция распределения $F_T(t)$.

При отсутствии информации о распределении случайной величины наиболее логичным представляется использование либо статистики $J_n(g_{1/2})$ (в этом случае предполагается, что величина T имеет вырожденное распределение в середине интервала, то есть $T = 1/2$ п.н.), либо статистики $J_n(g^{(u)})$ (предполагается равномерное распределение случайной величины T на отрезке $[0,\ 1]$). Доказывается, что статистики $J_n(g^{(u)})$ и $J_n(g_{1/2})$ приводят к функционалам $|Z_n(1/2)|$ и $\left|\int_0^1 Z_n(t)dt\right|$ соответственно. Отметим, что последний функционал — это функционал $J_n^{|1|}$, введенный равенством (3) при $r = 1$.

В результате проведенного исследования нами отобраны следующие функционалы:

$J_n^\infty = \sup_{t\in[0;\ 1]} |Z_n(t)|$ — супремум модуля эмпирического моста;

$J_n^{|r|} = \left|\int_0^1 (Z_n(t))^r dt\right|^{1/r}$, $r = 1,\ 2,\ \ldots$, — интегральные функционалы, совпадающие при четных r с L_r-нормами, а при нечетных оказывающиеся более предпочтительными в силу теоремы 1.1;

$J_n(g_{1/2}) = |Z_n(1/2)|$ — оптимальная статистика для случая, когда разладка происходит в середине интервала.

Для этих функционалов найдем приближенные бахадуровские наклоны.

Теорема 1.5 *Приближенные бахадуровские наклоны для функционалов J_n^∞, $J_n^{|r|}$, $J_n(g_{1/2})$ равны соответственно*

$$c_\infty(\theta) = 4{M_\theta}^2, \qquad c_r(\theta) = \frac{2^{2-2/r} B^2\left(\frac{1}{r},\ \frac{1}{2}\right)}{r(r+2)^{1-2/r}(r+1)^{2/r}} {M_\theta}^2,$$

$$c_{g_{1/2}}(\theta) = \left(\min\left\{\frac{1}{T};\ \frac{1}{1-T}\right\}\right)^2 {M_\theta}^2.$$

Здесь $B(\cdot,\cdot)$ — бета-функция.

Так как $c_\infty(\theta) > c_r(\theta)$ для любого $r \geq 1$, и $c_\infty(\theta) \geq c_{g_{1/2}}(\theta)$ для любого $T \in (0,\ 1)$, то наилучшим является критерий, использующий L_∞-норму. Он состоит в следующем: основная гипотеза отвергается в том и только том случае, когда

$$\sup_{t\in[0;\ 1]} |Z_n(t)| \geq C.$$

Если верна основная гипотеза Θ_0, то указанная статистика $J_n^\infty = \sup_{t\in[0;\ 1]} |Z_n(t)|$ сходится по распределению к случайной величине, имеющей распределение Колмогорова. Поэтому критерий имеет уровень ε, если $K(C) = 1 - \varepsilon$, где $K(x)$ — функция распределения Колмогорова. Для каждого фиксированного временного ряда достигнутый уровень значимости основной гипотезы равен $\varepsilon^* = 1 - K(J_n^\infty)$.

Чем меньше реально достигнутый уровень значимости ε^*, тем больше оснований отвергнуть основную гипотезу. В частности, основная гипотеза отвергается на уровне ε, если $\varepsilon^* \leq \varepsilon$.

Этот критерий будет использоваться в главе 4 для исследования однородности текстов на естественном языке.

Установление того, что этот критерий — наилучший во введенном классе, является главным результатом главы 1.

В **главе 2** изучена математическая модель временного ряда с долговременной зависимостью элементов — фрактальное броуновское движение. По определению, *фрактальное броуновское движение* с параметром Херста H, $0 < H \leq 1$ — это гауссовский процесс $X_H(t)$, $t \geq 0$, с нулевым математическим ожиданием и корреляционной функцией

$$\mathbf{E}X_H(t)X_H(s) = \frac{\sigma^2}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right). \tag{5}$$

Стандартное фрактальное броуновское движение $B_H(t)$ характеризуется условиями $B_H(0) = 0,\ \ \sigma = 1$.

Последовательность приращений $X_1, \ldots, X_n$ фрактального броуновского движения на единичных интервалах времени называется *фрактальным гауссовским шумом:*

$$X_i = B_H(i) - B_H(i-1).$$

В параграфе 2.2 рассмотрены существующие методы оценивания параметра Херста.

Если известно, что двумерный нормальный вектор имеет нулевой вектор математического ожидания (является центрированным), то существует взаимно однозначное соответствие между коэффициентом корреляции его компонент и вероятностью того, что эти компоненты имеют разный знак. Это соответствие служит основой для построения знаковых процедур оценивания параметра Херста, различные модификации которых составляют содержание параграфов 2.3 и 2.4.

Оценка элементарным знаковым методом определяется равенством

$$\tilde{H} = \frac{1}{2} + \frac{1}{2}\log_2\left(1 + \cos(\pi\nu)\right),$$

где ν — частота перемены знака, подсчитанная по выборке:

$$\nu = \frac{1}{n-1}\sum_{i=1}^{n-1} \mathbf{I}\{X_i X_{i+1} < 0\}.$$

Согласно теореме 2.1, оценка элементарным знаковым методом параметра $H \in (0,\ 1)$ сильно состоятельна.

В качестве более точного метода оценивания, учитывающего корреляции, возникающие при центрировании исходных данных, предлагается следующий *центрированный знаковый метод*, использующий конструкцию *фрактального броуновского моста* $B_H^0(t) = B_H(t) - tB_H(1)$.

Будем обозначать $X_i^* = X_i - \overline{X}$.

Обозначим $\tilde{p}_1^*$ — частоту перемены знака последовательностью $X_1^*,\ \ldots,\ X_n^*$.

Оценка $\widetilde{H}$ модифицированным знаковым методом определяется как решение уравнения

$$\tilde{p}_1^* = \frac{1}{\pi}\left(\arccos\left(2^{2\widetilde{H}-1} - 1\right) - \frac{2^{\widetilde{H}-1}}{\sqrt{1 - 2^{2\widetilde{H}-2}}} n^{2\widetilde{H}-2}\right).$$

В силу монотонности арккосинуса это уравнение при достаточно больших n имеет единственный корень на интервале $(0,\ 1)$ и может быть решено методом дихотомии.

Изложенный центрированный метод знаков полезен в силу того обстоятельства, что он позволяет работать с данными, имеющими произвольное математическое ожидание. Он позволяет в значительной мере компенсировать систематическую погрешность, возникающую при центрировании данных.

Для уменьшения дисперсии оценки предложены модифицированный знаковый метод и бинарный знаковый метод.

Обозначим через $S_{k,j}$ сумму k случайных величин, начиная с номера $j+1$. Обозначим $J_{k,j} = \mathbf{I}\{S_{k,j} \cdot S_{k,j+k} < 0\}$ — индикатор того, что соседние суммы k случайных величин имеют разные знаки, $L_k = \sum_{j=0}^{n-2k} J_{k,j}$ — число перемен знака соседними суммами k слагаемых.

Мы будем рассматривать статистики $V_n = V_n(K) = \sum_{k=1}^{K} L_k$. Это общее число перемен знака соседними суммами k слагаемых, $1 \le k \le K$. Здесь K — наибольшее допустимое число слагаемых в сумме, $K \le n/2$.

В качестве оценки вероятности перемены знака предлагается частота перемены знака, подсчитанная на основании статистики V_n:

$$p_n^*(K) = \frac{V_n(K)}{\sum_{k=1}^{K}(n-2k+1)} = \frac{V_n(K)}{K(n-K)}.$$

Для последующих вычислений будет полезно следующее обозначение: $G_{k,k',l}$ — ковариация индикаторов перемены знака парами блоков из k и k' слагаемых, середины которых смещены на l друг относительно друга. Эта ковариация вычисляется при условии, что $H = 1/2$.

Отметим, что по определению $G_{k,k',l} = G_{k',k,l}$. В теореме 2.4 вычислены коэффициенты $G_{k,k',l}$ в случае, когда $k' \le k$.

Обозначим

$$\sigma_K = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{|l|<\min\{k,\ k'\}} G(k,\ k',\ l)},$$

где $G(k,\ k',\ l)$ вычислены в теореме 2.4.

В следствии 2.1 доказана сходимость при условии $H = 1/2$ последовательности случайных величин $\frac{\sqrt{n}(p_n^*(K)-1/2)}{\sigma_K}$ по распределению к стандартному нормальному закону при $n \to \infty$.

Оценка модифицированным знаковым методом $\tilde{H}_n(K)$, вычисляемая на основании статистики $p_n^*(K)$, имеет вид

$$\tilde{H}_n(K) = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p_n^*(K)))\right).$$

Согласно теореме об асимптотической нормальности, эта оценка является асимптотически нормальной с коэффициентом $b_K(H) = |H'(p)|\sigma_K(H)$. При $p = H = 1/2$ получаем

$$b_K = b_K(1/2) = \frac{\pi}{2\ln 2}\sigma_K.$$

В качестве альтернативы модифицированному знаковому методу рассмотрим следующий *бинарный знаковый метод* оценивания параметра Херста. Он оказывается более удачным, чем модифицированный знаковый метод (оценки этим методом имеют меньшую дисперсию) в силу того, что здесь устранены большие положительные корреляции слагаемых.

Для упрощения изложения будем предполагать, что объем данных n является целой степенью числа 2, т. е. $\log_2 n$ — целое число.

Рассмотрим статистику

$$U_n = \sum_{k=0}^{\log_2 n-1} \sum_{j=0}^{n2^{-k}-2} \mathbf{I}\{S_{2^k,2^k j} \cdot S_{2^k,2^k(j+1)} < 0\}.$$

В качестве оценки вероятности перемены знака используется частота перемены знака, подсчитанная на основании статистики U_n:

$$p_n^* = \frac{U_n}{\sum_{k=0}^{\log_2 n-1}(n2^{-k}-1)} = \frac{U_n}{2n - 2 - \log_2 n}.$$

Оценка бинарным знаковым методом H_n^*, вычисляемая на основании статистики p_n^*, имеет вид

$$H_n^* = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p_n^*))\right).$$

Теорема 2.7 *Если $H = 1/2$, то при $n \to \infty$ имеет место слабая сходимость последовательности случайных величин $\sqrt{n}(p_n^* - 1/2)$ к нормальному закону с нулевым математическим ожиданием и дисперсией*

$$\sigma^2 = \frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s}\left(\frac{1}{\pi}\arccos 2^{-s/2} - \frac{1}{2}\right)^2 \approx 0,1654.$$

Следующее утверждение позволяет строить статистический критерий проверки гипотезы об отсутствии фрактальности $H = 1/2$ против ее альтернативы $H \neq 1/2$.

Следствие 2.4 *Если $H = 1/2$, то при $n \to \infty$ имеет место слабая сходимость последовательности случайных величин $\sqrt{n}(H_n^* - 1/2)$ к нормальному закону с нулевым математическим ожиданием и дисперсией*

$$B^2 = \frac{\pi^2}{4\ln^2 2}\left(\frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s}\left(\frac{1}{\pi}\arccos 2^{-s/2} - \frac{1}{2}\right)^2\right) \approx 0,8494.$$

Таким образом, при $n = 1024$ получаем стандартное отклонение $B/32 \approx 0,02880$. Этот результат оказывается лучше, чем для оценок модифицированным знаковым методом, и существенно лучше, чем для оценок элементарным знаковым методом и центрированным знаковым методом. Сравнение дисперсий также показывает, что эта оценка лучше оценки методом дисперсии, оценки нормированного размаха.

Основным результатом главы 2 является разработка бинарного знакового метода и обоснование его преимуществ перед другими методами.

В **главе 3** строятся алгоритмы проверки гипотез, связанных с моделью фрактального гауссовского шума.

Для проверки нормальности приращений в параграфе 3.2 строится новый критерий проверки нормальности, основанный на модификации знакового метода. Отметим, что как правило проверку нормальности предлагают проводить при объеме выборки n не менее 8. Это связано с тем, что для отыскания критических уровней используется нормальное приближение для используемой статистики, а оно оказывается весьма неточным. Однако уже при $n = 2$ можно (с нулевой вероятностью ошибки) отвергнуть гипотезу о нормальности, если выборочные значения совпадают. Для $n > 2$ предлагается построить критерий, наследующий это полезное свойство, но позволяющий также (с достаточно малой вероятностью ошибки) отвергать гипотезу о нормальности в некоторых случаях, когда все выборочные значения различны.

Обозначим $R_n = \max_{1\le i,j\le n} |X_i - X_j|$ — наибольшее расстояние между элементами (размах) выборки; $L_n = \min_{1\le i<j\le n} |X_i - X_j|$ — наименьшее расстояние между элементами выборки; $d_n = R_n/L_n$ — их отношение. Будем полагать $d_n = +\infty$ при $L_n = 0$.

Согласно основной гипотезе, элементы выборки $X_1, \ldots, X_n$ имеют нормальное распределение. Критерий отвергает основную гипотезу, если $d_n \ge C$, где $0 < C < \infty$.

Если выполнена основная гипотеза, то $d_2 = 1$ п.н. Рассмотрим статистику d_3.

$$d_3 = \frac{\max\{|X_1 - X_2|, |X_1 - X_3|, |X_2 - X_3|\}}{\min\{|X_1 - X_2|, |X_1 - X_3|, |X_2 - X_3|\}}.$$

В теореме 3.1 доказывается, что в случае выборки объема 3 любая статистика, не определенная при $R_3 = 0$, симметричная относительно элементов выборки и инвариантная относительно преобразований сдвига и масштаба, является функцией от d_3.

В следующих двух теоремах вычислены функции распределения статистик d_3 и d_4 в предположении, что основная гипотеза о нормальности верна.

Теорема 3.2 *Для всех* $x \geq 2$ *выполнено*

$$F_{d_3}(x) = 1 - \frac{6}{\pi}\mathrm{arcctg}\frac{2x-1}{\sqrt{3}}.$$

Теорема 3.3 *При* $x \geq 3$

$$F_{d_4}(x) = \frac{24}{\pi}\mathrm{arctg}\left(\mathrm{tg}\frac{a}{4}\sqrt{\mathrm{tg}\left(\frac{b}{2}+\frac{a}{4}\right)\mathrm{tg}\left(\frac{b}{2}-\frac{a}{4}\right)}\ \right),$$

где

$$a = \arccos\frac{x^2+6x-7}{3x^2-6x+11},$$

$$b = \arccos\frac{x^2+x-2}{\sqrt{3x^2-6x+11}\sqrt{x^2-2x+2}}.$$

Отметим, что осуществить вычисление распределения d_5 не удается ввиду отсутствия общей формулы, выражающей объем тетраэдра в пространстве постоянной кривизны через длины его ребер.

Для $n = 3,\ 4,\ 5$ выполнено сравнение с критерием Шапиро—Уилка методами математического моделирования. Значения квантилей уровня 0,05 для критерия Шапиро—Уилка исправлены по результатам моделирования 10^6 выборок. Сравнение критериев для различных альтернатив приводит к выводу о том, что R/L—критерий оказывается полезным при $n = 4,\ 5$ в случае существенной асимметрии распределения при альтернативной гипотезе (в частности, для логнормального распределения при $\sigma \geq 10$). При $n = 3$ мощности критериев совпадают, как и следует из теоремы 3.1.

В параграфе 3.3 на основании бинарного знакового метода разрабатывается и исследуется знаковый критерий для проверки основной гипотезы об отсутствии зависимости против гипотезы о соответствии модели фрактального броуновского движения.

В связи с тем, что не все гипотезы допускают аналитическую проверку, в параграфе 3.4 разработаны две процедуры моделирования фрактального броуновского движения: точный, но трудоемкий алгоритм разложения корреляционной матрицы на нижнюю и верхнюю треугольные; асимптотический метод — алгоритм скользящего среднего, основанный на соответствующих теоремах сходимости. Найдено, скольких слагаемых скользящего среднего достаточно для обеспечения заданной точности.

В параграфе 3.5 строится критерий выявления разладки фрактального броуновского движения, использующий результаты моделирования. Критерий основан на вычислении реально достигаемого уровня значимости по формуле

$$\varepsilon^*_{ij} = 2(1 - \Phi(|\tilde{H}_i - \tilde{H}_j - \tilde{a}_{ij}|/\tilde{\sigma}_{ij})),$$

где $\tilde{H}_i$, $\tilde{H}_j$ — оценки параметра H, полученные разными методами, а $\tilde{a}_{ij}$, $\tilde{\sigma}_{ij}$ — математические ожидания и среднеквадратические отклонения их разностей, вычисленные на основании моделирования для конкретного числа наблюдений. Предполагается асимптотическая нормальность разностей оценок, которая проверяется соответствующим статистическим критерием по результатам моделирования.

В параграфе 3.6 строятся критерии однородности двух фрактальных броуновских движений и их обобщений. Критерии основаны на статистике

$$J = \frac{\cdot \sum_{i=1}^{n} X_i}{\sum_{i=n+1}^{n+m} X_i}.$$

Последовательность X_i является фрактальным гауссовским шумом или его обобщением. Рассматриваемая статистика имеет распределение отношения зависимых нормальных случайных величин.

Первым рассматривается случай, когда X_i является фрактальным гауссовским шумом. В лемме 3.3 отыскиваются числовые характеристики числителя и знаменателя дроби $J = A/B$.

В теореме 3.5 строится интегральное представление для плотности распределения случайной величины J в этом случае.

В теореме 3.6 плотность распределения представляется в более явном виде через функцию Лапласа, что существенно ускоряет вычисления.

В параграфе 3.7 процедуры моделирования, оценивания параметров и проверки гипотез распространяются на последовательности, получаемые покоординатным преобразованием фрактального гауссовского шума для приведения в соответствие с симметричным устойчивым негауссовским законом распределения.

Моделируется стационарная случайная последовательность, элементы которой имеют симметричное абсолютно непрерывное устойчивое распределение. Совместное распределение элементов последовательности определяется тремя параметрами: параметром Херста, регулирующим характер зависимости; параметром устойчивого закона, определяющим скорость сходимости плотности распределения к нулю с ростом аргумента; масштабным параметром. При этом случайные величины являются зависимыми, и возможно регулировать характер зависимости приращений по аналогии с показателем Херста фрактального броуновского движения. Частными случаями таких последовательностей являются фрактальный гауссовский шум и последовательность приращений процесса Леви.

Преобразование приращений к автомодельному закону производится по формуле

$$Y_i = {F_\alpha}^{-1}(\Phi(X_i)),$$

где Φ — функция распределения стандартного нормального закона, F_α — функция распределения симметричного автомодельного закона. Параметр α может принимать значения из полуинтервала $(1,\ 2]$.

Оценивание параметров методом максимального правдоподобия по выборке $\mathbf{Y} = (Y_1, \ldots, Y_n)$ состоит в максимизации плотности совместного распределения

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}\sigma^n\sqrt{\det R}} \exp(-\frac{1}{2\sigma^2}\mathbf{X}^T R^{-1}\mathbf{X}),$$

где $\mathbf{X} = (X_1, \ldots, X_n)$,

$$X_i = \Phi^{-1}(F_\alpha(Y_i)).$$

Максимизацию нужно проводить по совокупности параметров α, H, σ. Однако такая процедура оказывается вычислительно сложной, и для выборок большого объема ее реализовать не удается. Поэтому используется

приближенный алгоритм: сначала отыскивается оценка параметра α в начальном приближении, затем она уточняется. Доказывается сильная состоятельность оценки. Затем производится нормализация значений и отыскиваются оценки параметров σ и H.

Нахождение оценки максимального правдоподобия и оценки бинарным знаковым методом осуществляется после этого преобразования так, как это было описано в соответствующих параграфах главы 2.

Сравниваются дисперсии оценок и строятся критерии для проверки основной гипотезы $H = 1/2$ против альтернативы в этой модели.

Главным результатом главы 3 является построение и обоснование статистических критериев проверки нормальности, зависимости приращений, разладки фрактального гауссовского шума и его обобщений.

Интернет можно рассматривать как систему, структурными элементами которой являются тексты. Возникает задача анализа этих элементов. Тексты могут иметь внутреннюю структуру, в частности, быть составленными из разнородных частей, что выявляется методами обнаружения разладки. Как будет показано, модель выборки не является вполне адекватной для моделирования текстов, и для однородных текстов используется модель фрактального гауссовского шума, а для разнородных — модель разладки фрактального гауссовского шума. Для анализа текстов используется математический аппарат обнаружения разладки в модели выборки, обнаружения зависимости элементов (отличия коэффициента Херста от $1/2$ в модели фрактального гауссовского шума), статистические критерии обнаружения разладки фрактального гауссовского шума, то есть изменения его параметров. Таким образом, в **главе 4** для статистического анализа текстов применяется весь математический аппарат, разработанный в предыдущих главах.

В параграфе 4.2 изучаются статистики текста на естественном языке. Текст рассматривается как последовательность слов $X_1, \ldots, X_n$. Здесь величины X_i, $i = 1, \ldots, n$, принимают значения в множестве $\mathcal{S}$, которое мы будем называть словником языка. Предполагается измеримость статистик относительно сигма-алгебры, порожденной событиями $\{X_i = X_j\}$, $1 \le i < j \le n$.

В частности, рассматривается статистика $N^{(n)}$ — число различных слов в тексте.

Предполагается, что $X_1, \ldots, X_n$ независимы и одинаково распределены с распределением $\mathrm{P}\{X_1 = i\} = p_i$, $i = 1, 2, \ldots$.

Согласно лемме 4.1,

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{\infty}(1 - (1 - p_i)^n). \tag{6}$$

Рассматриваются следующие однопараматрические модели:
1) $p_i = C(\alpha)i^{-\alpha}$, $i = 1, 2, \ldots$; $\alpha > 1$ (распределение Мандельброта);
2) $p_i = S(M)i^{-1}$, $i = 1, \ldots, M$; M — целое (распределение Ципфа);
3) $p_i = p(1-p)^{i-1}$, $i = 1, 2, \ldots$; $0 < p < 1$ (геометрическое распределение).
Здесь $C(\alpha)$, $S(M)$ — нормирующие константы, α, M, p — параметры.

Оценки параметров получаются подстановкой в (6) наблюдаемого значения числа разных слов вместо математического ожидания $\mathbf{E}N^{(n)}$. В теореме 4.1 доказана монотонность по параметру в формуле (6) для всех трех

моделей, что гарантирует однозначность оценки. В теореме 4.2 доказана состоятельность оценки в моделях 1 и 3.

В модели 2 состоятельность оценки $\tilde{M}_n$ очевидна. Теорема 4.3 доказывает несколько более общий факт для схемы серий: пусть $M = M_n$, $n \to \infty$, $\varepsilon > 0$. Если $(1+\varepsilon)M_n \ln^2 M_n \leq n$, то оценка $\tilde{M}_n$ состоятельна, то есть

$$\tilde{M}_n - M_n \to_p 0.$$

Совместно с М. Г. Чебуниным автором доказана функциональная центральная предельная теорема 4.4 для числа разных слов в модели более общей, чем модель 1.

Если вероятностная модель адекватна реальным данным, а оценка параметра в данной модели состоятельна, то оценка параметра с ростом объема выборки должна сходиться по вероятности к константе — истинному значению параметра. Так как для рассматриваемых моделей состоятельность доказана соответствующими теоремами, то отсутствие такой сходимости свидетельствует о неадекватности модели.

В параграфе 4.3 анализ однородности текста проводится методами, разработанными в главе 1.

Для анализа однородности текста необходимо определить способ, с помощью которого тексту сопоставляется последовательность чисел. Нами была разработана программа, которая сопоставляет тексту последовательность индикаторов вхождения слов текста в словарь авторского инварианта (служебных слов).

Было взято для исследования 25 текстов. Выбранные тексты исследовались поодиночке и в попарных комбинациях, полученных приписыванием одного текста к другому. Получилось $25 \times 24 = 600$ попарных комбинаций текстов, в том числе $3! + 9 \times 2 = 24$ комбинаций текстов одного автора и 576 комбинаций текстов разных авторов.

Для этих текстов были построены эмпирические мосты Z_n с помощью словаря авторского инварианта. Затем вычислялись максимальные по модулю отклонения эмпирического моста $J_n^\infty = \sup_{t\in[0;\,1]} |Z_n(t)|$, и отыскивался достигаемый уровень значимости $\varepsilon^* = 1 - K(J_n^\infty)$, где K — функция распределения Колмогорова.

Задача состоит в том, чтобы научиться различать одно произведение одного автора от склейки двух произведений разных авторов.

Для этого нужно выбрать уровень значимости ε и соответствующее предельное значение C.

Для произведения одного автора значения J_n^∞ не превосходят $3,4758$, что соответствует достигаемому уровню значимости $\varepsilon^* = 6,42 \cdot 10^{-11}$. Низкий достигаемый уровень значимости говорит о неполной адекватности модели выборки для описания появлений слов из выбранного словаря в тексте.

Проведен анализ 576 попарных комбинаций текстов разных авторов. Отметим, что здесь отклонение эмпирического моста от нуля может принимать большие значения, вплоть до 13.

При больших значениях J_n^∞ программа очень точно отыскивает границу между текстами. Так, текст «Аэлита»+«Мастер и Маргарита» делится точкой экстремума эмпирического моста на два фрагмента, первый из которых содержит лишь несколько строк из первой главы «Мастера и Маргариты».

Используются байесовкий (с равными априорными вероятностями гипотез) и минимаксный подходы к определению критического значения. Такой же анализ проводится для собраний сочинений и их фрагментов.

Анализ, проведенный в параграфе 4.3, показывает, что временные ряды, построенные по произведениям одного автора, собраниям сочинений или их фрагментам, не вполне соответствуют модели выборки — достигаемые уровни значимости часто близки к нулю. В параграфе 4.4 в качестве объяснения этого явления предлагается модель фрактального шума, которая исследуется на адекватность методами третьей главы.

Обозначим через $\tilde{H}_1$, $\tilde{H}_2$, $\tilde{H}_3$ оценки параметра методами нормированного размаха, дисперсии, знаков.

Анализ собраний сочинений и их фрагментов показывает, что уровни значимости гипотезы о фрактальном гауссовском шуме достаточно высоки за исключением случаев, когда текст существенно неоднороден (собрание сочинений Шолохова и его фрагменты 1 и 2). При этом лучше всего диагностируют разладку разности оценок $\tilde{H}_2 - \tilde{H}_3$ или $\tilde{H}_1 - \tilde{H}_3$, в которых одна из оценок отвечает за глобальные, а другая за локальные характеристики. В качестве критического уровня ε можно взять любое число от 0,01 до 0,1, при этом не будет ни одной ошибки обнаружения разладки.

Гипотеза об адекватности модели выборки не выдерживает проверки, достигаемые уровни значимости оказываются очень низкими, за единственным исключением («Детство» Л. Толстого, данные, содержащие после масштабирования всего 651 значение).

Итак, адекватной моделью процесса, описывающего появление служебных слов в тексте, для одного автора является фрактальный шум, для склейки текстов двух авторов — разладка фрактального шума. В качестве алгоритма обнаружения разладки предлагаются алгоритмы, основанные на разностях оценок $\tilde{H}_2 - \tilde{H}_3$ или $\tilde{H}_1 - \tilde{H}_3$.

Основным результатом главы 4 является обоснование использования фрактального шума в качестве модели текста и построение процедур статистического тестирования однородности текста.

В **главе 5** разрабатываются методы обнаружения разладки в регрессионных моделях.

Параграф 5.1 имеет вводный характер. В параграфе 5.2 изучается регрессия на порядковые статистики, доказываются утверждения, позволяющие построить сотоятельные критерии соответствия данных этой регрессионной модели.

В параграфах 5.3–4 изучается обнаружение разладки в регрессионных моделях с циклическим трендом. Предполагается, что $\{\varepsilon_i,\ i \geq 1\}$ — последовательность независимых одинаково распределенных случайных величин с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Пусть $\{g_j(\cdot),\ 1 \leq j \leq m\}$ — набор регрессионных функций, определенных на [0, 1]. Зададим треугольный массив $\{Y_{ni},\ 1 \leq i \leq n,\ n \geq 1\}$ зависимых переменных следующим образом:

$$Y_{ni} = \sum_{j=1}^{m} \theta_j g_j(i/n) + \varepsilon_i.$$

Итак, в этой модели все время наблюдений сжато на отрезок от нуля до единицы, и наблюдения отстоят друг от друга на равные интервалы

времени. В матричной записи

$$\mathbf{Y}_n = \mathbf{X}_n\theta + \varepsilon_n,$$

где $\mathbf{X}_n$ — матрица регрессора размерности $n \times m$, в которой на j-м месте в i-й строке стоит элемент $g_j(i/n)$; θ — вектор-столбец длины m; ε_n и $\mathbf{Y}_n$ — вектор-столбцы длины n.

Обозначим через $\widehat{\theta}_n$ оценку Гаусса—Маркова векторного параметра θ. Она равна

$$\widehat{\theta}_n = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{Y}_n,$$

если обратная матрица существует.

Если регрессионные функции интегрируемы с квадратом по Риману на $[0, 1]$, то существует предел

$$\lim_{n\to\infty} n^{-1}\mathbf{X}_n^T\mathbf{X}_n = G$$

— матрица, компоненты которой равны $\int_0^1 g_i(t)g_j(t)dt$.

Обозначим через $\mathbf{g}(\cdot)$ вектор-столбец регрессионных функций.

Частичные суммы регрессионных остатков обозначим через $\widehat{\Delta}_{nk} = \sum_{i=1}^k \widehat{\varepsilon}_{ni}$, где $\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$, $\widehat{\mathbf{Y}}_n = \mathbf{X}_n\widehat{\theta}_n$.

Будем полагать $\widehat{\Delta}_{n0} = 0$.

Рассмотрим эмпирический мост регрессионных остатков Z_n — случайную ломаную с узлами в точках

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_{nk} - \frac{k}{n}\widehat{\Delta}_{nn}}{\widetilde{\sigma}_n\sqrt{n}}\right),$$

где

$$\widetilde{\sigma}_n = \sqrt{\overline{\widehat{\varepsilon}^2} - \left(\overline{\widehat{\varepsilon}}\right)^2}.$$

В частном случае, когда регрессионные функции — это 1, $\cos 2\pi kt$, $\sin 2\pi kt$, $k \in M$, где M — некоторое конечное множество, модель принимает вид

$$Y_{ni} = a_0 + \sum_{k\in M} (a_k\cos(2\pi ki/n) + b_k\sin(2\pi ki/n)) + \varepsilon_i. \tag{7}$$

Рассматриваемая здесь регрессия — это модель с конечным числом взаимно ортогональных гармоник и аддитивным случайным шумом в дискретном времени. Эта модель известна в литературе как модель линейной регрессии с циклическим трендом. Изучение статистических критериев обнаружения разладки в этой модели предпринято в связи с исследованием колебаний строительных конструкций: ставится задача определения изменений прочностных характеристик конструкции по исследованию ее колебаний.

Предполагается, что моменты наблюдения равноотстоят друг от друга, и что период наблюдений состоит из целого числа периодов колебаний.

Модель (7) будем считать основной гипотезой, а в качестве альтернативной предполагается, что в некоторый момент времени значение a_0 в модели (7) заменяется на некоторое отличное от него значение b_0. При альтернативной гипотезе предполагается, что это изменение происходит только один раз за весь период наблюдений.

Задача состоит в том, чтобы построить класс статистических критериев, различающих основную и альтернативную гипотезы. После построения статистических критериев необходимо их сравнить и выбрать более мощный. Естественным подходом к обнаружению разладки, т. е. изменения параметров модели в процессе наблюдения, является анализ регрессионных остатков.

Простейшим критерием обнаружения разладки является критерий, основанный на сравнении средних значений остатков, подсчитанных по первой и второй половинам наблюдений. Ясно, что разность средних должна нормироваться среднеквадратическим отклонением или его выборочным аналогом. В результате получаем статистический критерий, близкий к критерию Стьюдента: большие по модулю значения нормированной разности средних свидетельствуют против основной гипотезы. Этот критерий, основанный на разности средних по первой и второй половине наблюдений, в терминах эмпирического моста основан на статистике $Z_n(1/2)$. Действительно, если число наблюдений n четно, то, так как сумма остатков регрессии в модели (7) равна нулю, разность между суммой первых $n/2$ и последних $n/2$ остатков равна $2\widehat{\Delta}_{n,n/2}$. При делении на $\widetilde{\sigma}_n\sqrt{n}$ получаем $2Z_n(1/2)$. При нечетном n логично использовать ту же статистику как обеспечивающую симметрию между первой и второй половинами наблюдений.

Обозначим $J_1 = Z_n^2(1/2)/\mathbf{D}B(1/2)$. При верной основной гипотезе статистика J_1 сходится слабо к распределению χ_1^2. Предложенный критерий можно обобщить следующим образом. Вместо одной точки $1/2$ взять d точек: $\frac{1}{d+1}, \ldots, \frac{d}{d+1}$, и рассмотреть декоррелированные и нормализованные (в соответствии с найденной корреляционной функцией) значения эмпирического моста Z_n в этих точках. Тогда сумма квадратов этих значений будет иметь в пределе хи-квадрат распределение с d степенями свободы в силу того, что сумма квадратов линейных функций от значений эмпирического процесса в фиксированных точках является непрерывным (в равномерной метрике) функционалом от эмпирического моста. И полученный таким образом критерий, и статистику, на которой он основан, будем обозначать J_d. Обозначим

$$\mathbf{z}_d = (z_{1,d}, \ldots, z_{d,d})^T = \left(Z_n\left(\frac{1}{d+1}\right), \ldots, Z_n\left(\frac{d}{d+1}\right)\right)^T.$$

Через C_d обозначим ковариационную матрицу вектора $\mathbf{b}_d = \left(B\left(\frac{1}{d+1}\right), \ldots, B\left(\frac{d}{d+1}\right)\right)^T$:

$$C_d = \mathbf{E}\mathbf{b}_d^T\mathbf{b}_d.$$

Статистика J_d вычисляется по формуле

$$J_d = \mathbf{z}_d^T C_d^{-1}\mathbf{z}_d.$$

Статистики критериев J_1, J_2, J_3 приведены в диссертации в явном виде (для последнего в случае, когда число наблюдений кратно четырем). На-

ряду с критериями J_1, J_2, J_3 будем использовать критерий, основанный на статистике $J = \sup_{t\in[0,1]} |Z_n(t)|$. Как показано в главе 1, этот критерий является асимптотически наиболее мощным в широком классе критериев в предположении, что $M = \emptyset$, т. е. в модели случайной выборки. В общем случае (при $M \neq \emptyset$) для предельного закона статистики нет аналитического описания, из общей теории известна лишь грубая асимптотика больших уклонений.

Смоделируем процесс (7) и сравним критерии, основанные на статистиках J_1, J_2, J_3 и J. Положим $n = 1024$, $M = \{4; 16\}$, $b_1 = b_2 = \sigma = 1$, $a_0 = a_1 = a_2 = 0$. В случае разладки параметру b_0 будем последовательно придавать значения от 0,2 до 1 с шагом 0,2. Момент разладки τ генерируется равномерно распределенным на целых числах от 1 до n.

Для того, чтобы скорректировать критерий J, а также проверить соответствие критических уровней для остальных критериев, было проанализировано поведение критериев при отсутствии разладки. Эмпирические уровни значимости были вычислены по результатам 20000 моделирований процесса. Для критериев J_1, J_2, J_3 эмпирические уровни значимости отличаются от выбранного теоретического уровня 0,05 не более чем на 0,002, а критерий J был скорректирован по результатам моделирования. В случае разладки предполагаем момент разладки τ равномерно распределенным на целых числах от 1 до n, а значению математического ожидания после разладки последовательно придаем значения от 0,2 до 1 с шагом 0,2. Вычисления проводятся с исправленным критерием J. Этот критерий по результатам моделирования при рассматриваемой альтернативной гипотезе является более мощным, чем критерии J_1, J_2, J_3, основанные на декорреляции и нормировке значений эмпирического моста в фиксированных точках. В то же время с ростом d мощность критерия J_d все ближе к мощности критерия J.

Главными результатами главы 5 являются построение и сравнение статистических критериев обнаружения разладки процесса гармонических колебаний со случайным шумом.

В **главе 6** разработанные в главе 1 методы обнаружения разладки применяются к медицинским и экономическим данным.

В параграфе 6.2 анализируется концентрация маркера M_i в крови пациента в зависимости от массы тела W_i. Предлагаются две регрессионных модели для объяснения зависимости. В первой модели $\ln M_i = \ln(a+b/W_i)+\varepsilon_i$, во второй $\ln M_i = a + b \ln W_i + \varepsilon_i$. Оценивание параметров в первой модели производится численно. Вычислением и анализом эмпирического моста показано преимущество второй модели.

В параграфе 6.3 анализируется цена квадратного метра квартиры в Новосибирске в зависимости от 697 параметров. Предлагаются 3 регрессионные модели. Анализ эмпирического моста показывает преимущество второй модели.

В параграфе 6.4 анализируется зависимость цены автомобиля «Тойота Королла» на вторичном рынке Новосибирска от года выпуска. Предложена модель $\ln Y_i = a + bX_i + \varepsilon_i$, $i = 1, \ldots, n$. Здесь X_i — год выпуска, Y_i — цена в рублях. Анализом эмпирического моста показана неприемлемость этой модели и предложена другая модель, с разрывом непрерывности при переходе от 1999 к 2000 году. Проверено, что эта модель согласуется с исходными данными.

Главным результатом главы 6 является демонстрация применения разработанных в диссертации методов для выбора наиболее подходящей вероятностной модели.

В **заключении** приведены основные результаты работы, составляющие научную новизну, теоретическую и практическую значимость. Обоснована достоверность результатов и сформулированы положения, выносимые на защиту.

Автор выражает глубокую признательность своим учителям Ю. Е. Хайкину, А. И. Саханенко, С. Г. Фоссу, В. А. Селезневу.

Публикации автора по теме диссертации

Индексируемые в базах цитирования

[1] Закревская Н. С., Ковалевский А. П. Однопараметрические вероятностные модели статистик текста // Сибирский журнал индустриальной математики. 2001. — Т. IV, N 2 (8). — С. 142–153. RSCI (ядро РИНЦ).

[2] Гусарова Г. В., Ковалевский А. П., Макаренко А. Г. Критерии наличия разладки // Сибирский журнал индустриальной математики. 2005. — Т. VIII, No. 4 (24). — С. 18–33. RSCI (ядро РИНЦ).

[3] Ковалевский А. П., Топчий В. А., Фосс С. Г. О стабильности системы обслуживания с континуально ветвящимися жидкостными пределами // Проблемы передачи информации. 2005. — Т 41, вып. 3. — С. 76–104. RSCI (ядро РИНЦ).

Kovalevskii A. P., Topchii V. A., Foss S. G. On the Stability of a Queueing System with Uncountably Branching Fluid Limits // Problems of Information Transmission. 2005. — V. 41, Issue 3. — P. 254–279. SCOPUS.

DOI 10.1007/s11122-005-0030-6

[4] Ковалевский А. П. Модифицированный знаковый метод тестирования фрактальности гауссовского шума // Проблемы передачи информации, 2008. — Т. 44, вып. 1. — С. 45–58. RSCI (ядро РИНЦ).

Kovalevskii A. P. Modified Sign Method for Testing the Fractality of Gaussian Noise // Problems of Information Transmission, 2008. — Vol. 44, No. 1. — P. 40–52. SCOPUS.

DOI 10.1134/S0032946008010043

[5] Ковалевский А.П., Костин В.С., Хиценко В.Е. Моделирование и идентификация последовательности зависимых случайных величин с симметричным устойчивым распределением // Сибирский журнал индустриальной математики, Том 13, 2010. — N 4 (44). — С. 25–37. RSCI (ядро РИНЦ).

[6] Аркашов Н.С., Ковалевский А.П. Вероятностная модель цен на квартиры // Сибирский журнал индустриальной математики, Том 15, 2012. — N 2 (50). — С. 11–20. RSCI (ядро РИНЦ).

[7] Ковалевский А. П., Шаталин Е.В. Асимптотика сумм остатков однопараметрической линейной регрессии, построенной по порядковым статистикам // Теория вероятностей и ее применения, Т. 59, N 3. — 2014. — С. 452–467. RSCI (ядро РИНЦ).

DOI 10.4213/tvp4579

Kovalevskii A. P., Shatalin E. V. Asymptotics of Sums of Residuals of One-Parameter Linear Regression on Order Statistics // Theory of Probability and Its Applications, Vol. 59, No. 3. — 2015. — P. 375–387. WoS, SCOPUS.

DOI 10.1137/S0040585X97T987193

[8] Kovalevskii A. P., Shatalin E. V. A limit process for a sequence of partial sums of residuals of a simple regression on order statistics with Markov-modulated noise // Probability and Mathematical Statistics, Vol. 36.1. — 2016. — С. 113–120. SCOPUS.

[9] Chebunin M., Kovalevskii A. Functional central limit theorems for certain statistics in an infinite urn scheme // Statistics and Probability Letters, V. 119. — 2016. — С. 344–348. WoS, SCOPUS.

DOI 10.1016/j.spl.2016.08.019

[10] Philonenko P., Postovalov S., Kovalevskii A. The limit test statistic distribution of the maximum value test for right-censored data // Journal of Statistical Computation and Simulation. 2016. — Vol. 86, iss. 17. — P. 3482–3494. WoS, SCOPUS.

DOI 10.1080/00949655.2016.1164703

[11] Ковалевский А. П. Тестирование нормальности очень малых выборок // Сибирские электронные математические известия, Т. 14. — 2017. — С. 1207–1214. WoS, SCOPUS.

DOI 10.17377/semi.2017.14.102

Другие статьи

[12] Филиппова Т.А., Ковалевский А.П., Русина Н.О. Основные вопросы маркетинга и менеджмента в энергетике // Научный вестник НГТУ, 1995. — N 1. — С. 161–169.

[13] Кувшинова М. А., Ковалевский А. П., Асланова И. В. Моделирование показателей энергопроизводства в системе поддержки принятия управленческих решений // Сборник научных трудов НГТУ, 2000. — No. 4 (21). — С. 133–138.

[14] Kovalevskii A. Dependence of increments in time series via large deviations // Proceedings of the 7th Korea-Russia International Symposium on Science and Technology. Ulsan, Korea, 2003. — P. 262–267.

[15] Закревская Н.С., Ковалевский А.П., Селезнева Л.В. Процесс Сопа // Научный вестник НГТУ. 2004. — N 3. — С. 13–19.

[16] Закревская Н.С., Ковалевский А.П. Алгоритм идентификации фрактального броуновского движения по разности оценок // Сборник научных трудов НГТУ, 2004. — N 2 (36). — С. 29–36.

[17] Ковалевский А. П. Применение принципа инвариантности к анализу однородности текста // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 195–204.

[18] Обухова О.О., Трунов А.Н., Ковалевский А.П. и др. Динамика продукции интерферона у больных герпетической инфекцией на фоне иммунокоррекции // Вестник новых медицинских технологий, 2008. — Т. XV, N 2. С. 141–143.

[19] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Оценка помехоустойчивости инвариантной системы связи при когерентном приеме // Электросвязь, 2009. – N 8. — С. 48–50.

[20] Алгазин Е.И., Ковалевский А.П., Левин Д.Н. Оценка помехоустойчивости системы обработки информации, инвариантной к мультипликативной помехе // Радиотехника, 2009. — N 6. — С. 28–31.

[21] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом при наличии аддитивной стационарной гауссовской помехи с корреляционной функцией общего вида // Вестник СибГАУ, вып. 1 (22), 2009. — С. 32–35.

[22] Алгазин Е.И., Касаткина Е.Г., Ковалевский А.П., Малинкин В.Б. Помехоустойчивость инвариантной системы передачи информации, основанной на когерентном приеме и при наличии слабых корреляционных связей // Вестник СибГАУ, вып. 2 (23), 2009. — С. 55–58.

[23] Алгазин Е.И., Ковалевский А.П., Касаткина Е.Г., Малинкин В.Б. Инвариантная когерентная система при комплексном воздействии помех // Вестник Тамбовского государственного технического университета, Т. 15, No. 2, 2009. — С. 295–302.

[24] Алгазин Е.И., Ковалевский А.П., Касаткина Е.Г., Малинкин В.Б. Инвариантная система при наличии аддитивной стационарной гауссовской помехи с корреляционной функцией общего вида и собственных шумов генераторного оборудования // Омский научный вестник, серия «Приборы, машины и технологии», є 2 (80), 2009. — С. 223–227.

[25] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом с последующей нелинейной обработкой // Вестник СибГАУ, вып. 3 (24), 2009. — С. 20–23.

[26] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Инвариантная система при нелинейной обработке сигналов // Омский научный вестник, серия «Приборы, машины и технологии», N 3 (83), 2009. — С. 272–275.

[27] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом с последующей нелинейной обработкой при наличии слабой корреляции // Вестник СибГАУ, вып. 4 (25), 2009. — С. 96–98.

[28] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Инвариантная система при нелинейной обработке сигналов и наличии слабой корреляции // Омский научный вестник, серия «Приборы, машины и технологии», N 1 (87), 2010. — С. 202–205.

[29] Алгазин Е. И., Ковалевский А. П., Малинкин В. Б. Способы повышения помехоустойчивости системы обработки информации, инвариантной к мультипликативной помехе // Радиотехника, 2010. — N 1. — С. 44–47.

[30] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Вопросы реализации оптимальной инвариантной системы передачи информации // Материалы X международной конференции «Актуальные проблемы электронного приборостроения», Том 4, 2010. — С. 123–125.

[31] Алгазин Е. И., Ковалевский А. П. Помехоустойчивость инвариантной системы при нелинейной обработке сигналов // В сб.: Современные проблемы радиоэлектроники. Красноярск, СФУ, 2011. — С. 505–509.

[32] A Posterior Change-Point Analysis in Application to the Dynamics of Enteric Infections and Water Turbidity in Ural Region of Russia. Kovalevsky A., Gubarev V., Loktev V. et al. // In: 22nd Annual Conference of The International Environmetrics Society, Book of Abstracts, Hyderabad, India, January 1–6, 2012. — P. 74.

[33] Kovalevskii A. A regression model for prices of second-hand cars // Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control. Novosibirsk, 2013. — P. 124–128.

[34] Ковалевский А. П. Сравнение статистических критериев разладки модели с циклическим трендом // Обозрение прикладной и промышленной математики, Т. 20, вып. 4, 2013. — С. 552–553.

[35] Шаталин Е. В., Ковалевский А. П. Асимптотика эмпирического моста в линейных регрессионных моделях, построенных по порядковым статистикам // Обозрение прикладной и промышленной математики, Т. 20, вып. 4, 2013. — С. 573–574.

[36] Шаталин Е.В., Ковалевский А.П. Асимптотика эмпирического моста в линейных регрессионных моделях, построенных по порядковым статистикам // Материалы XIV всероссийского симпозиума по прикладной и промышленной математике (осенняя сессия), Великий Новгород, — 2013. — С. 573–574.

[37] Ковалевский А.П. Статистические критерии обнаружения разладки регрессии с циклическим трендом // Научный вестник НГТУ. — 2013. — N 3 (52). — С. 55-62.

[38] Kovalevskiy A., Shatalin E. Limit processes for sequences of partial sums of residuals of regressions against order statistics with Markov-modulated noise // Conference program and abstract book of 11th International conference on ordered statistical data, Bedlewo(Poland). — 2014. — P. 37–38.

[39] Ковалевский А. П., Шахраманьян А. М. Анализ дефектов строительных конструкций методом эмпирического моста // Научный вестник НГТУ. — 2014. — N 3 (56). — С. 171–180.

[40] Ковалевский А. П., Шаталин Е.В. Выбор регрессионной модели зависимости массы тела от роста с помощью эмпирического моста // Вестник Томского государственного университета. Математика и механика, No.5(37). — 2015. — С. 35–47. (РИНЦ).

[41] Ковалевский А. П. Оценивание параметра закона Ципфа-Мандельброта по последовательности количеств разных элементов выборки // Обозрение прикладной и промышленной математики, 2017. — Т. 24, вып. 4. — С. 348–349.

[42] Kovalevskii A. P. Asymptotics of an empirical bridge of a regression on concomitants // 13 International conference on ordered statistical data (OSD 2018): conference program and abstract book, Spain, Cadiz, 22–25 May 2018. — 2018. — P. 29–30.

Монография

[43] Малинкин В.Б., Алгазин Е.И., Ковалевский А.П. Инвариантные системы связи. — Красноярск, 2010. — 202 с.

Авторские свидетельства

[44] Алгазин Е. И., Ковалевский А. П., Малинкин В. Б. Инвариантная система передачи информации по каналам с переменными параметрами. Патент на полезную модель N 85280. Зарегистрировано в Государственном реестре полезных моделей Российской Федерации 27 июля 2009 г.

[45] Свидетельство на программу для ЭВМ 2012660948. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании линейного детектора / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618953; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[46] Свидетельство на программу для ЭВМ 2012660949. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании поднесущей / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618954; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[47] Свидетельство на программу для ЭВМ 2012660950. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании синхронного детектора / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618955; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[48] Патент N 2014121189/08. Российская Федерация. МПК H03D3/00. Способ фазовой обработки сигналов / Алгазин Е. И., Ковалевский А. П.; заявитель и патентообладатель Новосиб. гос. техн. ун-т; заявл. 26.05.2014; опубл. 10.06.2016.

Ковалевский Артем Павлович

СТАТИСТИЧЕСКИЕ КРИТЕРИИ
АПОСТЕРИОРНОГО ОБНАРУЖЕНИЯ
РАЗЛАДКИ ВРЕМЕННЫХ РЯДОВ
И ИХ ПРИМЕНЕНИЯ

Автореферат диссертации на соискание ученой степени
доктора физико-математических наук

Подписано в печать 20.11.2018. Формат 60 × 84 1/16.
Бумага офсетная. Тираж 150 экз.
Печ. л. 2. Заказ N 1169.

Отпечатано в типографии
Новосибирского государственного технического университета
630073, г. Новосибирск, пр. К. Маркса, 20